J Int Comp & He Inf.
Vol. 1, No. 1, March 2020

pISSN: 2715-6923, eISSN: 2721-9186
https://jurnal.unimus.ac.id/index.php/jichi

Journal of Intelligent Computing and Health Informatics

Naïve Bayes Algorithm for Classification of Student Major's Specialization

Research Article

Astia Weni Syaputri[1,2*], Erno Irwandi[1,2], Mustakim[1,2]

[1]*Information System Department, Faculty of Science and Technology, UIN Sultan Syarif Kasim Riau, Pekanbaru 28293, Indonesia*
[2]*Puzle Research Data Technology (Predatech), Faculty of Science and Technology, UIN Sultan Syarif Kasim Riau, Pekanbaru 28293, Indonesia*

[*] *astiaweni.syaputri@students.uin-suska.ac.id (coresponden author)*
erno.irwandi@students.uin-suska.ac.id
mustakim@uin-suska.ac.id

Article history:
Received: 5 Jan 2020
Accepted: 2 Feb 2020
Available online: 31 Mar 2020

ABSTRACT

Majors are important in determining student specialization. If there is an error in the direction of the student, it will certainly affect the education of subsequent students. In SMA Negeri 1 Kampar Timur, there are two majors, namely Natural Sciences and Social Sciences. To determine these majors, it is necessary to reference the average value of student grades from semester 3 to semester 5 which includes the average value of Islamic religious education, Indonesian, Citizenship Education, English, Natural Sciences, Social Sciences, and Mathematics. Naive Beyes algorithm is an algorithm that can be used in classifying majors found in SMA Negeri 1 Kampar Timur. To determine the classification of majors in SMA Negeri 1 Kampar Timur, training data and test data are used, respectively at 70% and 30% . This data will be tested for accuracy using a confusion matrix, and produces a fairly high accuracy of 96.19%. With this high accuracy, the Naive Bayes algorithm is very suitable to be used in determining the direction of students in SMA Negeri 1 Kampar Timur.

Keywords : Confusion matrix, classification, naive bayes, student majors.

1. INTRODUCTION

The education is a process of guidance carried out consciously by educators of the process of physical and spiritual development of students, with the aim of forming a superior personality and can have sufficient meaning in (Baker, 1989). However, if education is not in accordance with what is desired by students will be fatal for the sustainability of education. Majoring is one of the educational processes desired by students. As a result what happens if students are wrong in determining the majors, there will be reluctance to learn and a decline in quality and academic achievement (Sofanudin, 2017).

In the 2013 curriculum, specialization of students in SMA Negeri 1 Kampar Timur was carried out based on the average grade of junior high school report cards in semester 3 through semester 5 and the majors desired by students (Liliana Swastina, 2013). Academics need to be careful and need special handling to determine the direction of students who are directed and right on target. Majors in secondary schools differ from other schools depending on the school. These majors include Natural Sciences (IPA) and Social Sciences (IPS) (Naparin, 2016). SMA Negeri 1 Kampar Timur is a high school that is in great demand by prospective students in the Kampar area, but there are still many students who turned out to be wrong in choosing a major. Of the 312 10th grade students, according to the grades they got during the two semesters there were about 45% of students who turned out to be wrong in choosing majors. This of course will cause the student's grades to fall and will result in the selection of higher education later.

In this study, a brief review of the use of the Naïve Bayes Classifier algorithm and its use in information on decision-making for majors. This research concentrates on specific problems that arise in applying the model to textual data. The research concludes with some thoughts on an interesting research direction for the Naive Bayes algorithm. Research conducted by (Alfa Saleh & Nasari, 2018) in his study entitled the classification of the Naïve Bayes method in data mining to determine the concentration of students in MAS BAO 2 Medan found that the Data Mining process using the Naïve Bayes method helped in obtaining information from the classification of student concentrations. Based on student academic data used as training data, the Naive Bayes method successfully classifies 109 student data from 120 tested data. So thus the Naive Bayes method was successful in predicting the concentration of students with an accuracy percentage of 90.8333% (Hastuti, 2012).

The results of research conducted by Husni Naparin (2016) regarding the classification of specialization of high school students from the initial stage to testing, and the results of the comparison can be concluded that the model formed with the Naive Bayes algorithm itself has a very good accuracy that is equal to 99.47% in classifying the status of students' specialization High school (A Saleh, 2015). Research conducted by (Alfa Saleh & Nasari, 2018) obtained test results from 100 student data with 90% accuracy. In this study, optimization of the method used previously by applying the Unsupervised Discretization technique will transform numerical / continuous criteria into categorical criteria and eliminate one criterion that is considered not to affect the accuracy of the test results, thereby increasing the accuracy of the classification results. From 120 student data tested, it is evident that the results of the classification of the application of the unsupervised discretization technique to the naive bayes method rose from 90% to 92.8% (Naparin, 2016). (Bisri, 2015) also conducted research on the classification of majors in Kesatrian High School and this method was able to assist the school in determining majors in Kesatrian High School. From the results of the prediction experiment student majors using matlab with the Naive Bayes method, obtained an accuracy of 83.8798% with an error rate of 16.1202% (Alfa Saleh & Nasari, 2018).

(Khasanah, 2016) examines the comparison of the high school student majors process using the Naïve Bayes classification algorithm and the ZeroR algorithm. The classification method with the Naïve Bayes algorithm is an algorithm that has the highest accuracy value of 96.74%. While the ZeroR algorithm has the lowest accuracy value of 59.78% (Bisri, 2015). Research conducted by (Yusra et al., 2016) comparing Naive Bayes and KNN using WEKA. Testing the accuracy of the method in this study was done with the 10-fold cross validation test option and the evaluation of test data using a confusion matrix. From the research that has been done, the results obtained on one hundred final assignment data with the number of random classes, Naïve Bayes method produces better accuracy values, which is 87%. Tests on the K-Nearest Neighbor method produce an accuracy value of 84% with a value of k = 3, 85% with a value of k = 5, 86% with a value of k = 7 and 84% with a value of k = 9 (Khasanah, 2016). By (Kadafi, 2018) also conducted a study comparing 3 classification algorithms namely Naïve Bayes, KNN, and C.45. Based on data processing and analysis conducted it was found that the Naïve Bayes algorithm is the best algorithm compared to other algorithms, with an accuracy rate of 79.51% and AUC at a value of 0.861 (Yusra et al., 2016). Therefore the Naive Bayes Clasifier algorithm was chosen as the research method for the classification of majors. The research conducted is to process the registration data stored so far in the database so that it can be used to obtain information for decision making for the school in doing its work.

2. MATERIALS AND METHOD

The method used in this study is an experimental research method, which consists of (1) data collection, (2) initial data processing (3) the proposed model, (4) testing the model, (5) Evaluation and validation of the model.

Secondary data is data obtained indirectly sourced from

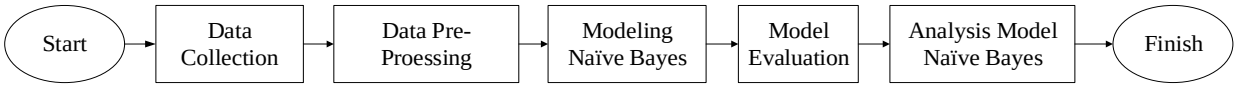

Fig. 1. Research method.

documentation, literature, books, journals and other information that has to do with the problem under study. Secondary data in this study are: books, journals about the Naïve Bayes Clasifier algorithm and data mining as well as data on new students of SMA N 1 Kampar Timur in 2018. While primary data is data obtained from research results. Primary data in this study are test result data using the Naïve Bayes Clasifier algorithm.

Data obtained from SMA Negeri 1 Kampar Timur are student data with the attribute Name, Address, School Origin, Grade 3 report cards up to semester 5 and the chosen majors. Other data to be processed is the average of each lesson value in semesters 3 through semester 5.

Data mining is a process that uses statistics, mathematics, artificial intelligence, and machine learning to extract and identify useful information and related knowledge from various large databases (Kadafi, 2018). So that data mining is a series of processes to explore the added value of a data set in the form of knowledge that has not been known manually (Nugroho, 2015). In data mining there are a series of methods used in solving problems including classification, clustering, association, prediction and estimation. One of the algorithms used in classification is Naïve Bayes. The flow of the Naive Bayes method is described in equations 1 to 7 (Bustami, 2014).

Naïve bayes theorem:

$$P(H|X) = \frac{P(X]H).P(H)}{P(H)} \quad (1)$$

To explain the Naive Bayes theorem, it is necessary to know that the classification process requires a number of clues to determine what class is suitable for the analyzed sample (Hasan, 2017).

$$P(C|F1 \dots FN) = \frac{P(C]) \, P(F1 \dots Fn|C)}{P(F1 \dots Fn)} \quad (2)$$

Where *C* Variable represents the class, while the variable of *F1 ... Fn* represents characteristic instructions needed to do the classification.

$$Posterior = \frac{(Prior \; x \; ;likehood)}{evidence} \quad (3)$$

The evidence values are always fixed for each class in one sample. The value of the posterior will then be compared with the value of the posterior grades of other classes to determine to which class a sample will be classified. Further elaboration of the Bayes formula is carried out by describing *(C / F1 ... FN)* using the multiplication rules as follows (Rosandy, 2016):

$$\begin{aligned} &P(C|F1 \dots FN) - P(C)P(F1 \dots Fn)C \\ &= P(C)P(F1 \; / C)P(F2, \dots Fn \; / C, F1) \\ &= P(C)P(F1|C)P(F2|C, F1)P(F3, \dots Fn|C, F1, F2) \\ &= P(C)P(F1|C)P(F2|C, F1)P(F3|C, F1, F2) \\ &P(F4, \dots Fn|C, F1, F2, F3) \\ &= P(C)P(F1|C)P(F2|C, F1)P(F3|C, F1, F2) \\ &\dots P(Fn \; / C, F1, F2, F3, \dots Fn - 1) \end{aligned} \quad (4)$$

It can be seen that the results of the elaboration cause more and more complexity of the condition factors that affect the probability value, which is almost impossible to analyze one by one (Hayuningtyas, 2017).

$$P(Pi|Fj) = \frac{P(Fi \cap Fj)}{P(Fj)} = \frac{P(Fi)P(Fj)}{P(Fj)} = P(Fj) \quad (5)$$

For $i \neq j$, so that

$$P(Pi|C, Fj) = P(Fi|C) \quad (6)$$

From the above equation it can be concluded that the assumption of naive independence makes opportunity conditions simple, so that calculations are possible. Furthermore, the translation of P (C | F1, ... Fn) can be simplified to:

$$\begin{aligned} &P(C|F1 \dots FN) - \\ &P(C)P(F1|C)P(F2|C)P(F3|C) \dots = \\ &P(C) \prod_{I=1}^{N} P \, (Fi|C) \end{aligned} \quad (7)$$

The equation above is a model of the Naive Bayes theorem which will then be used in the classification process. For classification with continuous data the Gauss Density formula is used:

$$P(Xi = xi|Y = Yi) = \frac{1}{\sqrt{2\sigma ij}} e^{\frac{(xi - \mu ij)^2}{2\sigma^2 ij}} \quad (8)$$

The model that has been developed in this research will be applied to the data of new students of SMA Negeri 1 Kampar Timur in 2018 through a simulation using the python programming language. The sample data consists of attributes of student names, majors, average scores of Islamic Education, average scores of Indonesian Language, average values of Civics Education, average English scores, average scores of Natural Sciences, average scores of IPS , and Mathematics grade point average. The amount of data sampled is 70% of the total data as training data and 30% of the total as test data.

3. RESULTS AND ANALYSIS

Majoring students of high school in SMA Negeri 1 Kumpar is important for the achievements that will be obtained by students. From the test results on the Naive Bayes algorithm then an analysis is carried out so that it can be concluded that the accuracy of the majors data is 96.19% based on the confusion matrix testing.

In the discussion in this study required some student value data which includes majors data taken by students, the average value of Islamic Education (PAI), the average value of Indonesian Language (B. IND), the average value of Civics Education (PKN) , the average value of English Language (B. ING), the average value of Natural Sciences, the average value of IPS, and the average value of MTK.data used can be seen in table 1.

Based on the data above will be classified using the Naive Bayes method into several classes, where this class is determined based on the majors in SMA Negeri 1 Kampar. The classes are (1) Natural Sciences and (2)

Table 1.Student grade dataset.

No.	Students names	Majors	Average of PAI Major's values	Average of B.Indo Major's values	Average of PKN Major's values	Average of B.Ingg Major's Values	Average of IPA Major's Values	Average of IPS Major's Values	Average of MTK Major's Values
1	Yohan Andrian Saputra	IPS	9,000	75,000	73,667	74,000	87,333	68,000	74,000
2	Winda Aulia Nisa	IPA	89,333	89,000	81,333	88,667	86,667	78,333	77,667
3	Nur Aviva Zahara	IPS	82,333	79,000	77,333	80,333	79,000	76,667	76,667
4	Mardhia	IPA	87,667	90,000	87,667	82,667	74,667	82,333	86,000
5	Amaliah Gusvita Sari	IPS	80,333	79,667	77,333	80,333	80,000	76,333	75,333
6	Syifa Fatimatul Zahra	IPA	86,667	87,333	86,333	84,667	80,333	76,333	88,000
7	Mus Ardicky Akbar	IPS	80,000	72,333	78,667	72,333	78,333	76,000	75,333
8	Deska Paturahman	IPA	88,667	76,333	73,000	71,333	81,000	81,333	83,333
9	Rifki Hidayat	IPS	81,667	79,667	79,333	79,333	82,000	74,000	75,000
10	Nurul Azizah	IPA	88,000	86,667	84,667	90,667	90,000	81,333	81,667
....	....	....	....	....	....	....	....	....	....
....	....	....	....	....	....	....	....	....	....
235	Erik Mahendra	IPS	73,667	69,000	74,333	69,333	76,333	74,667	69,000

Social Sciences.

Our experiments show that we produce new patterns, information and knowledge in the data mining process for the classification of student majors. From the data above we divide it into two, 70% training data are 246 data and 30% testing are 105 data. Training and testing data are used to determine students who are supposed to enter science or enter social studies.

From the process of calculating data mining using the Naive Bayes algorithm and the level of accuracy, new information is generated that is data mining calculations based on the test data of SMA Negeri 1 Kampar Timur students, showing students who choose IPS as many as 51 people, with a grade precision of 100%, and as many as 50 students who chose to enter science, and plus 4 people from students who chose IPS predicted to enter science with a class precision of 92.59%. From this result obtained an accuracy of 96.19%, with a true IPS class recall value of 92.73% and a true IPA class recall of 100%. A detailed explanation of the results of the classification can be shown in Table 2.

Table 2.Result of naïve bayes classification of studet grade dataset SMK Negeri 1 Kampar Timur.

	True IPS	True IPA	Class Precision
pred. IPS	51	0	100%
pred. IPA	4	50	92.73%
class recall	90.15%	100%	

In Table 3, the data analysis results using Naïve Bayes are shown, the data there are 4 students who were predicted to be wrong in choosing a major.

Table 3. Final results of the classification process.

Student's name	Majors	Predicted
Sindi Permata	IPS	IPA
Alma Yanti	IPS	IPS
Indah Islamiaty Taufiq	IPS	IPS
Abdullah	IPS	IPS
Septiani Putri	IPS	IPA
Fitri Ayu Andeska	IPS	IPS
Putri Amelia	IPS	IPS
Manja Dwi Permata	IPS	IPA
Alia Silvianur	IPS	IPS
Lia Oktari	IPS	IPA
…	…	…
Erik Mahendra	IPS	IPA

4. CONCLUSION

Based on our experiments and analysis, the conclusion is that the specialization of majors using the Naive Bayes algorithm results in a high degree of accuracy of the selection of majors as a planning for determining the next course. From the data mining calculation process using the Naive Bayes algorithm and the level of accuracy, a data mining calculation information is generated with a grade precision of 100%, and as many as 50 students who choose to enter science, and plus 4 of students who choose IPS predicted to enter science with a class precision of 92.59%. From this result obtained an accuracy of 96.19%, with a true IPS class recall value of 92.73% and a true IPA class recall of 100%. it can be concluded that the method of selecting majors using this algorithm is appropriate. Therefore, this research can be used for the purposes of selecting majors in the following years, because the resulting model is able to be predicted accurately with a

very small minimum error both in the selection of data and in terms of its classification.

REFERENCES

Baker, R. C. (1989). Nonlinear unstable systems. *International Journal of Control*, *23*(4), 123–145.

Bisri, M. H. (2015). Implementasi Algoritma Naïve Bayes untuk Memprediksi Penjurusan Siswa di SMA Kesatrian 1 Semarang. *Jurnal Informatika*, 1–7.

Bustami. (2014). Penerapan Algoritma Naive Bayes Untuk Mengklasifikasi Data Nasabah Asuransi. *TECHSI - Jurnal Teknik Informatika*, *8*(1), 127–146. https://doi.org/10.26555/jifo.v8i1.a2086

Hasan, M. (2017). Menggunakan Algoritma Naive Bayes Berbasis. *ILKOM Jurnal Ilmiah*, *9*(3), 317–324.

Hastuti, K. (2012). Analisis Komparasi Algoritma Klasifikasi Data Mining untuk Prediksi Mahasiswa Non-Aktif. *Seminar Nasional Teknologi Informasi & Komunikasi Terapan 2012 (Semantik 2012)*, *2*(1), 241–249.

Hayuningtyas, R. Y. (2017). Aplikasi Filtering of Spam Email Menggunakan Naïve Bayes. *IJCIT (Indonesian Journal on Computer and Information Technology)*, *2*(1), 53–60.

Kadafi, A. R. (2018). Perbandingan Algoritma Klasifikasi Untuk Penjurusan Siswa SMA. *Jurnal ELTIKOM*, *2*(2), 67–77. https://doi.org/10.31961/eltikom.v2i2.86

Khasanah, F. N. (2016). Klasifikasi Proses Penjurusan Siswa Tingkat SMA Menggunakan Data Mining. *Informatics for Educators and Professionals*, *1*(1), 65–69.

Liliana Swastina. (2013). Penerapan Algoritma C4.5 untuk Penentuan Jurusan Mahasiswa. *Jurnal Gema Aktualita*, *2*(1), 93–98.

Naparin, H. (2016). Klasifikasi Peminatan Siswa SMA Menggunakan Metode Naive Bayes. *Systemic: Information System and Informatics Journal*, *2*(1), 25–32. https://doi.org/10.29080/systemic.v2i1.104

Nugroho, Y. S. (2015). Klasifikasi dan Klastering Penjurusan Siswa SMA Negeri 3 Boyolali. *Khazanah Informatika: Jurnal Ilmu Komputer Dan Informatika*, *1*(1), 1. https://doi.org/10.23917/khif.v1i1.1175

Rosandy, T. (2016). Perbandingan Metode Naive Bayes Classifier dengan Metode Decision Tree (C4.5) untuk Menganalisa Kelancaran Pembiayaan (Study Kasus : KSPPS / BMT AL-FADHILA). *Jurnal Teknologi Informasi Magister Darmajaya*, *2*(01), 52–62.

Saleh, A. (2015). Klasifikasi Metode Naive Bayes Dalam Data Mining Untuk Menentukan Konsentrasi Siswa. *KeTIK*, 200–208.

Saleh, Alfa, & Nasari, F. (2018). Penggunaan Teknik Unsupervised Discretization pada Metode Naive Bayes dalam Menentukan Jurusan Siswa Madrasah Aliyah. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *5*(3), 353. https://doi.org/10.25126/jtiik.201853705

Sofanudin, A. (2017). Program studi sistem informasi fakultas teknik universitas nusantara pgri kediri tahun 2017. *Simki-Techsain*, *01*(03), 1–6.

Yusra, Olivita, D., & Vitriani, Y. (2016). Perbandingan Klasifikasi Tugas Akhir Mahasiswa Jurusan Teknik Informatika Menggunakan Metode Naïve Bayes Classifier dan K-Nearest Neighbor. *Sains, Teknologi Dan Industri*, *14*(1), 79–85. https://doi.org/10.1002/mame.201200226